Genetic Algorithm Operations Implementation on Traffic Anomaly Intrusion Detection using NSLKDD Dataset

Nagarajan Munusamy, L.Gnanaprasanambikai

Abstract*: *Intrusion Detection is one of the security tools in Network Security. Anomaly Intrusion Detection is a method of Intrusion Detection to find novel attacks by using decision rules. These decision rules need to be globally fit solution to find new attacks. Genetic Algorithm is a evolutionary search algorithm, which gives the globally fit solutions in the state space search. Genetic Algorithm process success depends on its fitness function. In this paper, a suitable fitness function is proposed to find global fit rules, for Traffic anomaly intrusion detection. For a better Intrusion Detection performance the discussion of proposed fitness function results with test data and comparison with existing fitness function are done and tabulated.

Index Terms*: *Fitness Function, Genetic Algorithm, GA-Operations, Anomaly Intrusion Detection.

I. INTRODUCTION

Network Security is very much important in the modern internet world. Security threats to the network increased lot. So various measures are used in organization to detect and prevent the security threats. The Intrusion Detection is one such measure to monitor the actions occurring in computer system or network and alarm the network administrator the intrusions. Intrusion Detection is classified into two categories based on the monitoring action they are Host Intrusion Detection System (HIDS) and Network Intrusion Detection System(NIDS). HIDS execute on individual host or devices in a network and monitors the packets for any intrusions. NIDS execute on devices at some strategic points in the network and monitors the traffic for any undesirable actions [1].

Intrusion Detection uses a set of rules for monitoring the actions in the network. Further, Intrusion Detection is classified into two categories based on the detection method, they are Signature(or Misuse) Intrusion Detection and Anomaly Intrusion Detection. Signature Based IDS detect intrusions based on observed data that matches with pre-defined description on the rules about the intrusion action. The advantage of Signature based IDS is that accurate intrusion detection with low false positive rate and dis-advantage is lack of detecting new intrusions. Anomaly Based IDS detect intrusions by detecting anomaly from observed data which deviates from normal behavior. The advantage of Anomaly IDS is that detects new intrusions and disadvantage generation of false alarms [1].

Soft Computing is a new research field which includes technologies for intelligent problem solving. The objective of Soft Computing is to exploit fault tolerance, partial truth and uncertainty problems by achieving tractability, robustness and low cost solutions [2]. Genetic Algorithm is one constituent of Soft Computing is used mimics biological evolution for finding optimal solution of a problem. The application of Genetic Algorithm in Intrusion Detection is used to find fittest rules for monitoring with less false alarms. To find the fitness of the rule, GA uses fitness function.

In this paper, we implement GA for Anomaly intrusion detection with a proposed fitness function which finds both false positive rate and false negative rate of a rule and finds the global Optimum Rules. The future work of [18] is implemented in this paper.

Paper Organization

We have brief introduction of Intrusion Detection. The rest of the paper is organized as follows. Section 2 presents paper related to this research work. Section 3 presents Overview of Genetic Algorithm. Section 4 presents Proposed Approach of Genetic Algorithm in IDS. Section 5 presents Experimental Results and Evaluation. Section 6 presents Conclusion. Section 7 presents future work.

II. RELATED WORK

Wei. Li used temporal and spatial information of network connections in deriving rules and used Genetic Algorithm approach to detect anomalous network intrusion. Weighted penalty sum model is used in GA as fitness function to show detection rate. But the paper doesn't show any experimental results [3].

R.H.Gong et.al., and A.A.Ojugo et.al., used both categorical and quantitative of network data for deriving the classification rules, and the used Genetic Algorithm approach to detect misuse network intrusion. Support Confidence Framework is used as fitness function in GA to show high detection rate. The paper showed the implementation method and experimental results. For the computational purpose, Whole training dataset is loaded into memory but this is infeasible and inefficient for large training dataset [4][5].

Revised Manuscript Received on June 11, 2019

Dr. Nagarajan Munusamy, HOD & Associate Professor, K.S.G College of Arts & Science, Bharathiar University, Coimbatore, Tamil Nadu.
L.Gnanaprasanambikai, Assistant Professor, Nehru Arts and Science College, Bharathiar University, Coimbatore, Tamil Nadu.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

S.S.Kandeeban et. al.,used penalty factor as the fitness function in Genetic Algorithm approach to detect anomalous behavior in intrusion detection. The paper presented software implementation of the proposed approach with high attack detection and low false positive rate[6].

A.Goyal et., al., V.M.Hashemi et.,al., B.Abdullah et., al., used GA approach has a common fitness function which is used for both misuse and anomaly intrusion detection. The fitness function considers on low false positive rate only and high detection rate. However false negative rate is not considered[7][8][9].

Priya U. Kadam et.,al., used GA approach for different types of attack in anomalous intrusion detection. The fitness function used in GA is based on chromosome strength. The paper presents the analysis of detection rate for different types of attack and specifies the attack which is suitable for this fitness function[10].

F.Alabsi et., al., used GA approach with reward-penalty fitness function to detect attacks. The fitness function gives reward to good chromosomes and penalty to bad chromosomes. The paper also presents a comparison between existing fitness function and proposed fitness function[11].

III. GENETIC ALGORITHM OVERVIEW

Genetic Algorithm is adaptive heuristic search Algorithm based on the evolutionary ideas of natural selection and genetics. It is a search technique to find approximate solutions to optimization and search problems. Genetic Algorithm begins with a set of solutions to a problem and called as initial population. These solutions are represented as chromosomes which is a abstract representation. Each chromosome is evaluated with a fitness function. Once the fitness function is properly defined and a set of Genetic algorithm operations are done until the termination condition becomes true. The termination condition is a convergence criterion which may be maximum number of generations, elapsed time, and no improvement in the fitness function of the chromosomes [12].

Pseudo code of Genetic Algorithm operations

```
Begin
Set of initial population
Evaluate each chromosome of a population with fitness function
Repeat until the termination condition with the following steps
    ✓ Select the chromosomes based on their fitness function
    ✓ Crossover the gene of the chromosomes with certain probability and produce new offspring
    ✓ Mutate the offspring with certain probability
    ✓ Replace the individuals of the old population with new population
End [13]
```

IV. PROPOSED APPROACH

We use NSL KDD Dataset for experimental purpose. NSLKDD dataset is de-facto dataset for anomaly intrusion detection. NSLKDD dataset is advanced version of KDDCUP99. The advantage of this dataset is no redundancy, no duplication of records and less complexity level. NSLKDD dataset consists of 20% training dataset, full training dataset of 125973 instances and testing dataset of 22544 instances [14]. For our experiment initially we do Data Preprocessing using PCA, dimensionality reduction method, for feature extraction and feature set construction for traffic anomaly intrusion detection[15]. Then we create decision rules from 20%training dataset using C4.5 decision tree which is implemented as J48 classifier in WEKA3.6 [18]. We could generate 31 decision rules with normal class [15] for anomaly intrusion detection

The proposed approach is classified into two stages they are: training stage and testing stage. In training stage we use full training dataset to train the decision rules and to find the global optimum rules. In the testing stage, the global optimum rules are used for intrusion detection.

The decision rules are constructed in the form of if-then rules. A rule in intrusion detection system consists of two parts one is condition (antecedent) and other is action (consequent). To evaluate intrusion detection following terminologies are followed [1]

True Negative (TN) - If condition and action both are true means the rule classifies normal data as normal.
False Positive (FP) - If condition is false and action is true means the rule classifies normal data as intrusion.
False Negative (FN) - If condition is true and action is false means the rule classifies intrusion data as normal
True Positive (TP) - If condition is false and action is false means the rule classifies intrusion data as intrusion
These rules are evaluated with the following proposed fitness function which is an unconstraint optimization fitness function

$$\text{True fitness} = \left(\frac{\text{TP}}{\text{TP}+\text{FN}}\right) + \left(\frac{\text{TN}}{\text{TN}+\text{FP}}\right) \tag{1}$$

$$\text{False Fitness} = \left(\frac{\text{FP}}{\text{TN}+\text{FP}}\right) + \left(\frac{\text{FN}}{\text{TP}+\text{FN}}\right) \tag{2}$$

$$\text{Fitness} = \text{True Fitness} - \text{False Fitness} \tag{3}$$

The Generating algorithm for searching global optimum rules performed as follows. The first step is initial population for the mating pool [13] is to be formed from the 31 decision rules (chromosomes), with the roulette wheel selection method to know the probability of occurrence [2]. We found only 29 decision rules are suitable for mating pool formation. Then the genetic algorithm operations (selection, crossover, mutation, replacement) are applied for rules in the mating pool with the training data loaded. Next the Genetic algorithm operations are executed with different iteration level with the proposed fitness function. The output of global optimum rules of

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

each iteration is used for intrusion detection in the testing dataset [8].

V. IMPLEMENTATION STEPS

A. Training stage

The implementation and experiments are done using MATLAB. The initial population for the mating pool are formed with 29 decision rules. The following Genetic Algorithm steps are performed for several iterations. In each iteration global optimum rules are selected, and tested for detection rate, false alarms. The proposed fitness function and existing fitness function are implemented for the following steps and the results are compared.

Step 1: Selection Technique

In this section two GA selection mechanisms are presented [17]. They are

a. Random Selection

The Random selection technique is performed by generating two random different parents in the population (each represents a chromosome). Then the genetic algorithm operations are applied to the randomly selected chromosomes.

b. Binary Tournament Selection

The following steps are performed twice for binary tournament selection

- ✓ Generating two random different parents (each represents chromosomes).
- ✓ Two randomly selected chromosomes fitness values are compared and the one with the better fitness value will go into the mating pool.

From the above two best randomly selected chromosomes are applied with genetic algorithmic operations.

Step 2: Crossover Technique

Crossover is merging the genetic information of two parents. We performed single-point crossover, which is simple and most often used. In this method, randomly a crossover point is chosen on both the parents. Gene is swapped between the two parents, on the randomly chosen crossover point. Two offspring are produced on applying the technique **[**17].

Parent 1	P1g1	P1g2	P1g3	P1g4
Parent 2	P2g1	P2g2	P2g3	P2g4

Crossover point⟶

Offspring1	P1g1	P1g2	P2g3	P2g4
Offspring2	P2g1	P2g2	P1g3	P1g4

From the above the table, two parents Parent1 and Parent2 namely P1 and P2 each with four genes namely, g1, g2, g3 and g4 respectively. On a random crossover point, the genes are recombined between the two parents and two offspring namely, Offspring1 and Offspring2 are produced. To our problem, chromosome represents the rule and gene represents the attributes in chromosome.

Step 3: Mutation Technique

Mutation is the change in the genetic material of the offspring produced due to erroneous crossover. We performed random mutation. In this mutation a new gene is produced from the randomly chosen gene domain [16].

Step 4: Replacement Technique

Replacement is the last stage in the breeding cycle. It decides which individuals to be replaced in the population to maintain the size of it. When two parents produce two offspring, total four individuals cannot return to the population, only two can be replaced with the current members of the population. We found best two individuals based on the fitness function and performed random replacement. In this replacement two best individuals replaces two randomly individuals of population inclusive of parents for selection. This method is suitable for population with small size [13].

Results and summary

We compared the results of the proposed approach with the existing fitness function. As per the discussion of various fitness function in section 2, the fitness function which is commonly used for misuse and anomaly intrusion detection of recent years is taken as existing function which also gives high detection rate and is a constraint optimization fitness function.

Existing Fitness Function

$$fitness = \frac{\mathrm{a}}{\mathrm{A}} - \frac{\mathrm{b}}{\mathrm{B}} \qquad (4)$$

Where

A=True Positive (TP) and False Negative (FN)
B=True Negative (TN) and False Positive (FP)
a=True Positive (TP)
b=False Positive (FN)

The detection rate of the existing fitness function is obtained by not considering false negative parameter rate. Our proposed approach considers all the four parameters TP, TN, FP, FN and favors the true value. The following two tables give the Global fitness value for various iterations of both existing and proposed fitness function in random selection and binary tournament selection.

Random Selection

TABLE I *Global Fitness value of Random Selection*

Iterations	Proposed fitness			Existing fitness		
	Global fitness	**Iteration no**	**No. of Fitness rule**	**Global fitness**	**Iteration no**	**No. of Fitness rule**
20	.9793	18	22	.9079	12	18
50	1.0306	49	12	.9454	14	13
75	1.1636	75	10	.9079	12	17
100	1.2703	100	4	.9914	94	2
150	1.3485	144	4	.9104	12	5

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Binary Tournament Selection

TABLE II *Global Fitness value of Binary Tournament Selection*

Iterations	Proposed fitness			Existing fitness		
	Global fitness	Iteration no	No. of Fitness rule	Global fitness	Iteration no	No. of Fitness rule
20	.9254	20	17	.9200	18	9
50	1.0783	50	13	.8943	7	20
75	1.1152	75	8	.8889	5	22
100	1.2362	100	3	.9023	97	4
150	1.2565	120	3	.8889	5	26

We could find the proposed fitness function value increases as iteration increases and is high compared to all existing fitness function. In comparison with random selection and binary tournament selection of the proposed fitness function we could find the random selection gives higher fitness value as iteration increases and gives high number of global fitness rules.

B. Testing Stage

In this stage testing dataset is loaded for evaluation of high detection rate, false alarms and accuracy which is calculated with the following formulas. These formulas are used for accessing Intrusion Detection performance.

$$Accuracy = \ \ TP + TN$$

$$\text{No. of errors} \ \ = FP + FN$$

$$\text{Detection Rate} \ = \left(\frac{\text{Accuracy}}{TP + TN + FP + FN}\right) * 100$$

$$\text{False Rate} = \frac{(\text{No. of Errors})}{(TP + TN + FP + FN)} * 100$$

TABLE III *Detection Rate of Existing and Proposed Fitness function*

	Iteration Number	Existing Fitness Function		Proposed Fitness Function	
		Detection Rate	False Rate	Detection Rate	False Rate
Random selection	20	84.5413	15.4587	86.7459	13.2451
	50	82.9711	17.0289	86.7016	13.2984
	75	84.5413	15.4587	87.0653	12.9347
	100	64.2388	35.7612	85.4108	14.5892
	150	81.6714	18.3286	85.4108	14.5892
Binary Tournament selection	20	73.5362	26.4638	77.1868	22.1832
	50	83.3526	16.6474	84.1155	15.8845
	75	79.2983	20.7107	80.6733	19.3267
	100	61.2713	38.7827	64.2255	35.7745
	150	79.2983	20.7107	83.3880	16.6120

Table 3 gives the Detection rate and False Rate of proposed fitness function and existing fitness function with random selection and tournament selection methods for various iterations. The Global fitness value obtained from the training stage is applied on the testing dataset and the detection results are taken.

In comparison of proposed fitness function and existing fitness function we could find proposed fitness function gives high detection with low false rate both in Random Selection and Binary Tournament Selection at all iteration level.

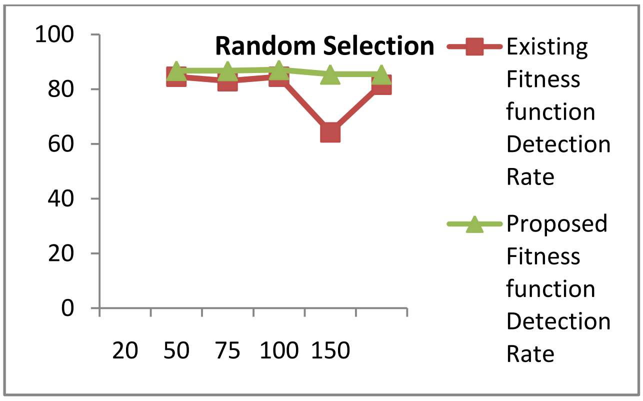

Figure 1: Existing and Proposed Fitness function Detection Rate for Random Selection

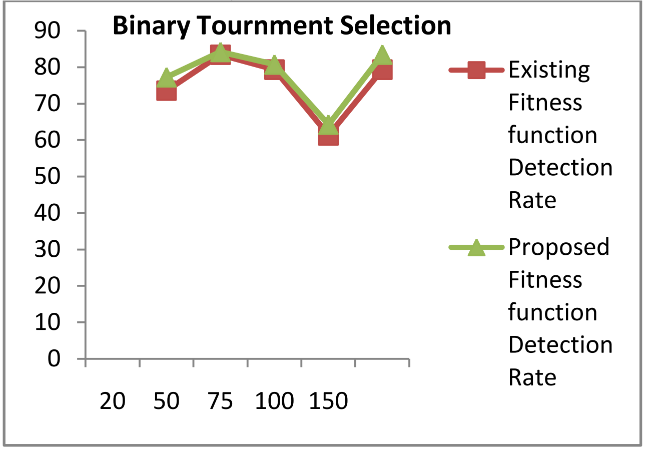

Figure 2: Existing and Proposed Fitness function Detection Rate for Binary Tournament Selection

In comparison of proposed fitness function with random selection and binary tournament selection, we could find random selection gives high detection rate and low false rate.

In the figure 3, we could find at all iteration level, the proposed fitness function for random function l the detection rate is approximately the same and for binary tournament selection the detection rate rise and fall.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

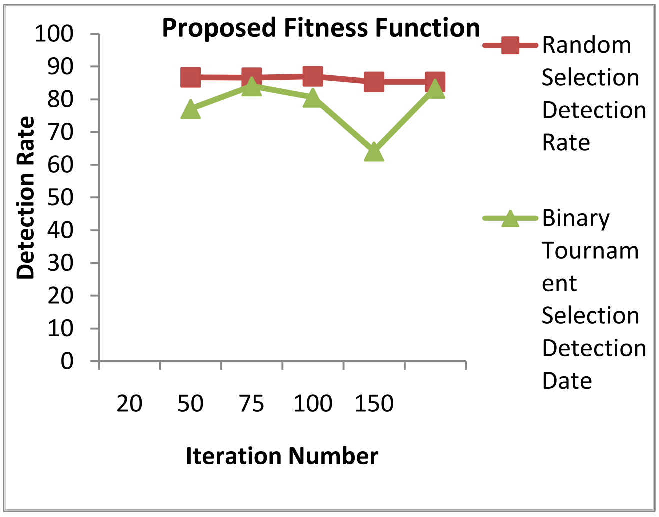

Figure 3: Proposed Fitness Function – Detection Rate Comparison Chart for Random Selection and Binary Tournament Selection.

VI. CONCLUSION

In this paper, we implemented Genetic Algorithm operations for traffic anomaly intrusion detection with a proposed fitness function. The drawback of existing fitness function and need for proposed fitness function is all specified. We compared the results of the proposed fitness function with the exiting fitness function at various generations and found proposed fitness function gives higher detection rate by considering both false negative and false positive parameters. We could find the random selection of Genetic Algorithm process is suitable for the proposed fitness function which gives better results. The implementation results are tabulated and results are discussed. In future the same work can be implemented in real time.

REFERENCES

1. K.G. Srinivasa, N.Pramod: gNIDS: rule-based network intrusion detection systems using genetic algorithms, International Journal of Intelligent Systems Technologies and Applications, Vol.11, Nos 3/4, pp: 252-266,2012.
2. S. Rajesekaran, G.A.Vijayalaksmi Pai: Neural Networks, Fuzzy logic and Genetic Algorithms Synthesis and Applications, PHI, India, ISBN-978-81-2023-2186-1, 2011.
3. Wei. Li: Using Genetic Algorithm for Network Intrusion Detection, SANS Institute, USA, 2004.
4. R.H.Gong , M.Zulkernine and P. Abolmaesumi: A software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection, Proceedings of the 6th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks, , Towson, Maryland, USA,pp:246-253, May 23-25, 2005.
5. A.A.Ojugo, A.O.Eboka, O.E.Okonto, R.E.Yoro, F.O.Aghware: Genetic Algorithm Rule-Based Intrusion Detection System, Journal of Emerging Trends in Computing and Information Science, Vol.3, No.8, August 2012.
6. S.S.Kandeeban, R.S.Rajesh; A Mutual Construction for IDS using GA, International Journal of Advanced Science and Technology, Vol.29, pp:1-8, April 2011.
7. A.Goyal and C.Kumar: Genetic Algorithm based Network Intrusion Detection System, from: http://www.cs.northernwestern.edu/~ago210/ganids/GANIDS.pdf
8. V.M.Hashemi, Z.Muda and W.Yassin: Improving Intrusion Detection Using Genetic Algorithm, Information Technology Journal 12(11):2167-2173, 2013.
9. B.Abdullah, I.Abd-alghafar, G.I. Salama, A.Abd-alhafez: Performance Evaluation of a Genetic Algorithm Based Approach to Network Intrusion Detection, 13th International Conference on Aerospace Sciences & Aviation Technology, ASAT-13, pp: 1-17, 2009.
10. Priya U. Kadam, P. P. Jadhav: An effective rule generation for Intrusion Detection System using Genetics Algorithm, International Journal of Science, Engineering and Technology Research, Volume 2, Issue 10, October 2013.
11. F.Alabsi and R.Naoum: Fitness Function for Genetic Algorithm used in Intrusion Detection System, International Journal of Applied Science and Technology, Vol.2, No.4, PP.129-134, April 2012.
12. P.G.Majeed, S.Kumar: Genetic Algorithms in IntrusionDetection Systems: A Survey, International Journal of Innovation and Applied Studies, ISSN 2028-9324 Vol. 5 No. 3, pp. 233-240 Mar. 2014.
13. S.N.Sivanandam, S.N.Deepa: Introduction to Genetic Algorithms, ISBN 978-3-540-73189-4 Springer, 2008.
14. S.Revathi, Dr.A.Malathi: A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection, International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 12, ISSN: 2278-0181, pp: 1848-1853, December – 2013.
15. R.Goel, A. Sardana, R.C. Joshi: Parallel Misuse and Anomaly Detection Model, International Journal of Network Security, Vol.14, No.4, pp: 211-222, July 2012.
16. Muhammad Tami Al-Hajri, M. A. Abido: Assessment of Genetic Algorithm Selection, Crossover and Mutation Techniques in ReactivePower Optimization, 2009 IEEE Congress on Evolutionary Computation (CEC 2009),1005-1011, 2009.
17. Jorge Magalhães-Mendes: A Comparative Study of Crossover Operators for Genetic Algorithms to Solve the Job Shop Scheduling Problem, WSEAS TRANSACTIONS on COMPUTERS,Vol 12, Issue 4, ,pp:164-173, April 2013.
18. L.Gnanaprasanambikai, Dr. Nagarajan Munusamy, “Data Preprocessing and Classification for Traffic Anomaly Intrusion Detection using NSLKDD Dataset”, Cybernetics and Information Technologies, Vol 18 No3, pp:110-119, Sep 2018.

AUTHORS PROFILE

Dr. Nagarajan Munusamy is currently Head and Associate Professor, Department of Multimedia and Web Technology, KSG College of Arts and Science, Coimbatore, Tamil Nadu, India . He has completed Ph.D in Computer Science in 2012. He has published many reputed International Journals and has an experience more than 15 years in the Industry and Academic. His research area includes Wireless Sensor Networks, Remote Sensing, Image Processing and Network Security.

L.Gnanaprasanambikai, received her Bachelor Degree in Computer Science from Bharathiar Univerisity. Completed her Post graduation in Computer Communication from Bharathiar University. Finished M.Phil in Computer Science from Bharathiar University. Currently Pursuing Ph.D. in Bharathiar Univeristy. Her research are includes Network Security, machine learning techniques.